ANALISA KOMPUTASI KEMUNCULAN DAN KEPUNAHAN KOSAKATA BAHASA INDONESIA BERDASARKAN CORPUS

Muhammad Fachrul Kurniawan[1], Faisal Rahutomo[2], Ridwan Rismanto[3]

[1,2,3] Teknik Informatika, Teknologi Informasi, Polteknik Negeri Malang
[1]fachrul.xiia1.26@gmail.com, [2]faisal.polinema@gmail.com,[3]ridwan@polinema.ac.id

Abstrak

Bahasa Indonesia merupakan bahasa yang biasa kita gunakan sehari-hari. Penelitian ini tertarik untuk mengungkap seberapa besar tingkat kepunahan dan kemunculan kosakata baru di dalam pengucapan sehari-hari karena juga terdapat bahasa baru yang sering digunakan oleh anak-anak muda jaman sekarang. Penelitian ini bertujuan untuk menganalisa hubungan antara kosakata asing maupun kosakata yang telah punah terhadap waktu sehingga kita dapat mengetahui penggunaan kosakata asing maupun kosakata yang punah dari waktu ke waktu. disini peneliti menggunakan data kata dari berita online untuk diteliti selama enam bulan dari berbagai situs berita online untuk menganalisa peningkatan maupun penurunan kosakata tersebut. Metode analisis yang digunakan adalah analisis regresi sederhana. Dari hasil penelitian ini, menunjukan bahwa terdapat tingkat penurunan kata asing yang digunakan pada berita online dari minggu ke minggu, sedangkan kepunahan bahasa Indonesia pada berita online cenderung mengalami peningkatan

Kata Kunci: kata punah, kata asing, analisa regresi

1. Pendahuluan

Bahasa Indonesia merupakan bahasa kesatuan yang dapat menyatukan seluruh rakyat Indonesia, namun seiring berjalannya waktu bahasa Indonesia baku mulai tersisih dengan bahasa-bahasa asing maupun bahasa dari daerah-daerah yang ada diseluruh Indonesia. Masyarakat lebih sering menggunakan bahasa dari daerah masing-masing dikarenakan pengucapannya mudah mereka ingat karena banyak yang menggunakan bahasa dari tiap-tiap daerah daripada menggunakan bahasa Indonesia yang baku. Bukan hanya bahasa daerah saja yang sering digunakan sebagai bahasa obrolan sehari-hari, bahasa asing juga sering digunakan untuk komunikasi kita sehari-hari. Sehingga membuat bahasa Indonesia menjadi asing bagi kita karena jarang diucapkan bahkan ada yang sampai punah karena tidak pernah diucapkan oleh kebanyakan orang.

Penelitian ini tertarik untuk mengungkap seberapa besar tingkat kepunahan dan kemunculan kosakata baru di dalam pengucapan sehari-hari karena juga terdapat bahasa baru yang sering digunakan oleh anak-anak muda jaman sekarang. Bahasa baru itu pulalah yang sering digunakan dalam obrolan sehari-hari, contohnya seper ti *baper*, *alay*, *lebay*, dan lain sebagainya. Selain bahasa baru juga terdapat berbagai macam bahasa Indonesia baku yang terdapat pada Kamus Besar Bahasa Indonesia (KBBI), namun jarang atau bahkan tidak pernah kita jumpai.

Maka dari itu penelitian ini mengusulkan suatu sistem komputasi yang memungkinkan kita bisa menganalisa kosakata bahasa Indonesia yang kita gunakan sehari-hari untuk menentukan kemungkinan bahasa yang jarang kita ucapkan atau bahkan tidak pernah terucapkan sama sekali. Selain itu sistem ini diharapkan juga dapat menemukan kata-kata yang tidak ada pada Kamus Besar Bahasa Indonesia namun sering digunakan untuk percakapan sehari-hari. Dalam proses analisa terkomputasi ini penulis menggunakan data dari berita *online*.

2. Landasan Teori

2.1. Operasi Teks

Operasi teks diperlukan untuk memperoleh kata-kata yang bermakna untuk mereprensentasikan sebuah dokumen. Operasi teks yang dilakukan disebut praproses dokumen yang meliputi tahapan-tahapan berikut:

1. Analisis leksikal teks dengan tujuan menghilangkan angka, tanda hubung, tanda baca, dan besar huruf.
2. Penghilangan *stopword* dengan tujuan memfilter kata yang memiliki nilai rendah untuk keperluan *retrieval*.
3. *Stemming* kata yang tersisa dengan tujuan menghilangkan imbuhan.
4. Pemilihan term untuk digunakan sebagai elemen pengindeksan.

2.2 *Tokenizing*

Istilah *token* didefinisikan sebagai unit terkecil dari sebuah teks atau dapat juga merupakan suatu

kumpulan dari *string alphanumerik*. Suatu unit terkecil yang digunakan disini adalah sebuah kata tunggal yang disebut juga sebagai *term*. Sebuah kata tunggal dapat berisi sekumpulan *string alphanumerik*. Di dalam proses ini terjadi pemotongan dokumen menjadi daftar kata yang berdiri sendiri sebelum dilakukan proses selanjutnya.

2.3 Analisa Regresi

Analisa Regresi adalah suatu metode analisis statistik yang digunakan untuk melihat pengaruh antara dua atau lebih variabel. Hubungan variabel tersebut bersifat fungsional yang diwujudkan dalam suatu model matematis. Pada analisis regresi, variabel dibedakan menjadi dua bagian, yaitu variabel respons (*response* variable) atau biasa juga disebut variabel bergantung (*dependent* variable) dan variabel *explanory* atau biasa disebut penduga (*predictor* variable) atau disebut juga variabel bebas (*independent* variabel).

Analisis regresi digunakan hampir pada semua bidang kehidupan, baik dalam bidang pertanian, ekonomi dan keuangan, industri dan ketenagakerjaan, sejarah, pemerintahan, ilmu lingkungan, dan sebagainya. Kegunaan analisis regresi diantaranya untuk mengetahui variabel-variabel kunci yang memiliki pengaruh terhadap suatu variabel bergantung, pemodelan, serta pendugaan (*estimation*), atau peramalan (*forecasting*).

Model persamaan regresi linear adalah seperti berikut ini:

$$\mathbf{Y = a + bX}$$

Dimana:
Y = Variabel *Response* atau Variabel Akibat (*Dependent*)
X = Variabel *Predictor* atau Variabel Faktor Penyebab (*Independent*)
a = konstanta
b = koefisien regresi (kemiringan); besaran *Response* yang ditimbulkan oleh *Predictor*.
Tahap Analisa Regresi:

1. Penentuan Tujuan
2. Identifikasikan Variabel Penyebab dan Akibat
3. Pengumpulan Data
4. Hitung a dan b berdasarkan rumus Regresi Linear Sederhana

$$\mathbf{a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}}$$

$$\mathbf{b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}}$$

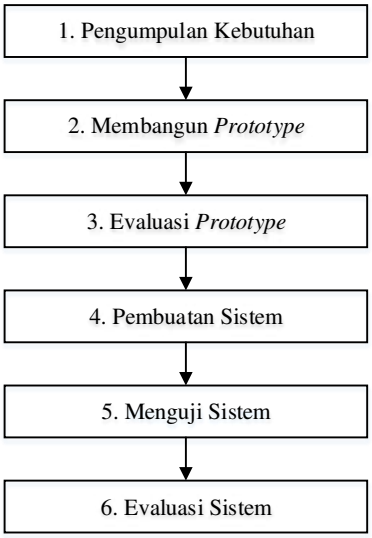
5. Buat Model Persamaan Regresi

$$\mathbf{Y = a + Bx}$$

3. Analisis dan Perancangan

Pada bagian ini dibahas metodologi yang digunakan peneliti dalam pembuatan Analisa Komputasi Kemunculan dan Kepunahan Kosakata Bahasa Indonesia Berdasarkan Corpus Berita Online. Metode penenlitian yang digunakan adalah metode *prototype*.

Proses kegiatan yang ada pada metode *prototyping* dapat dijelaskan pada Gambar 1 sebagai berikut.



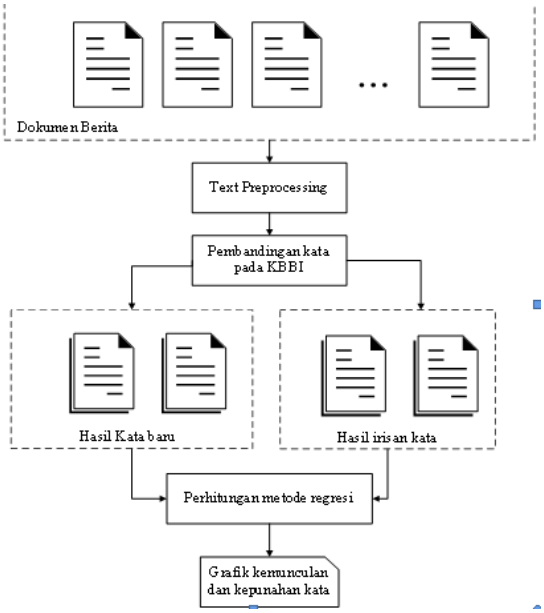
Gambar 1. Prototype Model

Berikut merupakan penjelasan dari setiap tahap *prototype*.

1. Pengumpulan Kebutuhan
   Pada tahap ini hal yang dilakukan adalah mendefinisikan kebutuhan perangkat, sistem ataupun data yang diperlukan untuk menganalisa kemunculan dan kepunahan kosakata bahasa Indonesia.
2. Membangun *Prototype*
   Pada tahap ini dimulai perancangan sementara pada sistem yang dibuat, sehingga *user* dapat melihat fungsi-fungsi dari program yang dibuat.
3. Evaluasi *Prototype*
   Mengevaluasi perancangan yang telah dibuat. Jika telah sesuai dengan yang diinginkan maka dilanjutkan ke tahap berikutnya. Jika belum sesuai maka mengulang pada tahap 1, 2, dan 3.
4. Pembuatan Sistem
   Proses pembuatan sistem yang sesuai dengan perancangan. Dari beberapa perancangan yang sudah dijelaskan pada tahap sebelumnya, maka pada tahap ini mulai mengimplememtasikan hasil perancangan tersebut ke dalam sistem.

5. Menguji Sitem
   Pengujian dilakukan dengan cara menghitung akurasi kebenaran dari pengelompokkan data berita.
6. Evaluasi Sistem
   Mengevaluasi dan dan memperbaiki sistem jika sistem yang dibuat belum berjalan sesuai yang diharapkan. Kemungkinan adanya *bug/error* pada tahap pengujian sistem, maka pada tahap ini akan diselesaikan.

Dibawah ini merupakan Gambaran umum dari system Analisa komputasi kemunculan dan kepunahan kosakata bahasa Indonesia :



3.1. Metode Pengumpulan Data

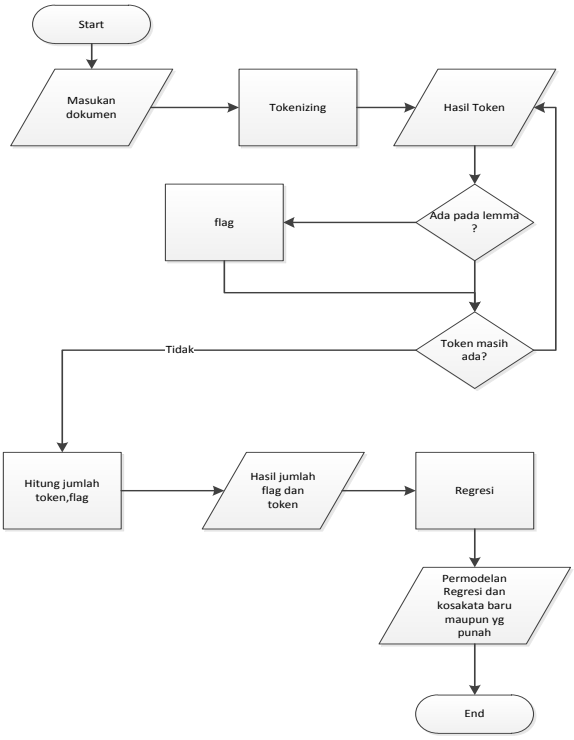
Data-data yang didapatkan untuk melakukan penelitian yaitu melalui hasil *crawling* data berita dari situs berita terkait. Data tersebut diperoleh per hari selama 6 (enam) bulan.

3.2. Perancangan Sistem

Perancangan sistem merupakan suatu proses desain sistem dalam penggambaran dan pembuatan sketsa *interface* aplikasi hingga perhitungan dari metode itu sendiri yakni *Regresi linear*. Rancangan ini sendiri akan terbagi menjadi tiga yakni perancangan perhitungan metode, perancangan proses dalam bentuk *flowchart,* dan perancangan *user interface* atau *mockup* dari aplikasi ini.

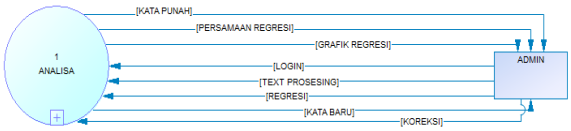
3.3. Flowchart System

Flowchart merupakan sebuah diagram yang digunakan untuk menjelaskan alur dari sebuah sistem dengan menggunakan simbol-simbol yang telah ditentukan dan saling terhubung.
Dibawah ini merupakan flowchart system :



3.4. Perancangan Data Flow Diagram (DFD)

DFD adalah sebuah teknis grafis yang menggambarkan aliran informasi dan transformasi yang diaplikasikan pada saat data bergerak dari *input* menjadi *output.*DFD dapat dipisahkan ke dalam level-level yang merepresentasikan aliran data yang lebih detail (Pressman, 2002).
Berikut merupakan rancangan DFD level 0 :



Gambar 2. DFD Level 0 (*Context Diagram*)

4. Uji Coba dan Pembahasan

Pada bab pengujian dan pembahasan ini akan dilakukan tahapan untuk menguji hasil dari implementasi sistem yang telah dilakukan.

4.1. Pengujian Sistem

Pengujian sistem ini dilakukan dengan cara menjalankan aplikasi secara detail pada setiap menu yang ada, dengan tujuan untuk mengetahui menu atau fitur mana yang sudah berfungsi dengan baik maupun yang tidak berfungsi sesuai dengan sebagaimana mestinya.

4.2. Pembahasan

Dari penelitian yang telah dilakukan. Disini akan dibahas mengenai analisa kosakata Bahasa Indonesia terhadap waktu.

Berikut ini merupakan bahasan dari analisa kata baru terhadap waktu :

Tabel Kemunculan kata baru Tiap Minggu

No	Waktu (X)	Kata Baru (Y)	X^2	Y^2	XY
1	1	26126	1	682567876	26126
2	2	33634	4	1131245956	67268
3	3	26281	9	690690961	78843
4	4	24407	16	595701649	97628
5	5	30849	25	951660801	154245
6	6	30583	36	935319889	183498
7	7	30313	49	918877969	212191
8	8	31312	64	980441344	250496
9	9	32136	81	1032722496	289224
10	10	31390	100	985332100	313900
11	11	30409	121	924707281	334499
12	12	30883	144	953759689	370596
13	13	28078	169	788374084	365014
14	14	30766	196	946546756	430724
15	15	29621	225	877403641	444315
16	16	28545	256	814817025	456720
17	17	30283	289	917060089	514811
18	18	29065	324	844774225	523170
19	19	29003	361	841174009	551057
20	20	29622	400	877462884	592440
21	21	28101	441	789666201	590121
22	22	29615	484	877048225	651530
23	23	28399	529	806503201	653177
24	24	29238	576	854860644	701712
25	25	31107	625	967645449	777675
26	26	25772	676	664195984	670072
27	27	22697	729	515153809	612819
Total (S)	378	788235	6930	23165714237	10913871

Dari table diatas dapat diketahui model persamaan regresinya yaitu

$$Y= 30231.66 -74.12 X$$

Pada persamaan diatas dapat diketahui bahwa kata asing cenderung menurun dari minggu ke minggu Berikut ini merupakan bahasan dari analisa kata punah terhadap waktu :

Tabel Kepunahan kosakata Tiap Minggu

No	Waktu (X)	Kata Punah (Y)	X^2	Y^2	XY
1	1	30161	1	909685921	30161
2	2	28930	4	836944900	57860
3	3	30034	9	902041156	90102
4	4	30292	16	917605264	121168
5	5	29405	25	864654025	147025
6	6	29344	36	861070336	176064
7	7	29475	49	868775625	206325
8	8	29412	64	865065744	235296
9	9	29486	81	869424196	265374
10	10	29476	100	868834576	294760
11	11	29611	121	876811321	325721
12	12	29522	144	871548484	354264
13	13	29752	169	885181504	386776
14	14	29341	196	860894281	410774
15	15	29607	225	876574449	444105
16	16	29615	256	877048225	473840
17	17	29482	289	869188324	501194
18	18	29646	324	878885316	533628
19	19	29658	361	879596964	563502
20	20	29553	400	873379809	591060
21	21	29858	441	891500164	627018
22	22	29523	484	871607529	649506
23	23	29794	529	887682436	685262
24	24	29844	576	890664336	716256
25	25	29407	625	864771649	735175
26	26	30157	676	909444649	784082
27	27	30671	729	940710241	828117
Total (S)	378	801056	6930	23769591424	11234415

Dari table diatas dapat diketahui model persamaan regresinya yaitu

$$Y= 29500.95 + 11.98 X$$

Pada persamaan diatas dapat diketahui bahwa kata punah dari minggu ke minggu cenderung meningkat

5. Kesimpulan

a. Pada Analisa kemunculan kosakata bahasa Indonesia dapat disimpulkan bahwa penggunaan kata asing dari waktu ke waktu semakin berkuarang namun setelah peneliti mencari penyebab mengapa terdapat kata asing ternyata kata asing pada berita online tersebut dipengaruhi kebanyakan dari nama produk, nama orang, nama perusahaan hingga banyak kata-kata yang dikarenakan salah pengetikan selain itu juga terdapat kata-kata dari bahasa yang memang bukan dari Indonesia. Namun walaupun demikian pertumbuhan kata asing pada berita online dari minggu ke mingu cenderung menurun.
b. Pada Analisa kepunahan kosakata bahasa Indonesia dari minggu ke minggu semakin meningkat hal ini dapat dilihat dari pembahasan untuk kata punah, yang dimaksud kata punah disini merupakan kata-kata bahasa Indonesia yang terdapat pada KBBI namun tidak terpakai pada dokumen berita online. dengan demikian dapat disimpulkan bahwa keragaman kata yang digunakan pada situs berita online dari minggu ke minggu cenderung menurun.

6. Saran

Berdasarkan dari penelitian dikarenakan keragaman bahasa Indonesia yang terdapat pada situs berita online semakin menurun maka diharapkan dari berbagai penyedia situs berita online untuk meningkatkan keragaman kata yang digunakan pada berita karena bila keragaman kata yang terdapat pada KBBI apabila tidak sering digunakan maka berangsur-angsur kata tersebut akan menjadi asing bahkan bagi bangsa kita sendiri.

Daftar Pustaka

Bambang Kurniawan, Syahril Efendi, Opim Salim Sitompul. 2012. *"Kasifikasi Konten Berita Dengan Metode Text Mining"* Jurnal Dunia Teknologi Informasi Vol. 1, No. 1, (2012) 14-19. Universitas Sumatera Utara.

Saraswati, N. S. (2011). *Text Mining Dengan Metode Naivde Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis*. Tesis Pada Universitas Udayana

http://documents.software.dell.com/statistics/textbook/text-mining

https://nahulinguistik.wordpress.com/2010/04/19/pergeseran-pemertahanan-dan-kepunahan-bahasa/
http://www.pengertianahli.com/2014/07/pengertian-regresi-apa-itu-regresi.html